01 - What is Multivariate Data?

Junvie Pailden

SIUE, F2017, Stat 589

August 23, 2017

Multivariate Methods

Multivariate data includes simultaneous measurements on many variables.

Most multivariate analysis involves analysis of measurements obtained with out actively controlling or manipulation any of the variables on which the measurements are made.

1. Data reduction or structural simplification
2. Sorting and grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and testing

Organization of Data

- n measurements on $p \geq 1$ number of variables
- $x_{jk} =$ measurement of the kth variable on the jth term
- Let $\mathbf{X}_{n \times p}$ matrix that contains the data consisting of all obs on all variables

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Descriptive Statistics

Sample mean, sample variance

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \quad s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{jk} - \bar{x}_k)^2, \quad k = 1, 2, \ldots, p$$

Sample covariance measures linear association between the ith and kth variables.

$$s_{ij} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k), \quad i, k = 1, 2, \ldots, p$$

Sample correlation coefficient is a standardized version of the sample covariance

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

Arrays of Basic Descriptive Statistics

Sample mean vector $(p \times 1)$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample variance covariance matrix

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample correlation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Intuition and pitfalls for correlation

Correlation = LINEAR association

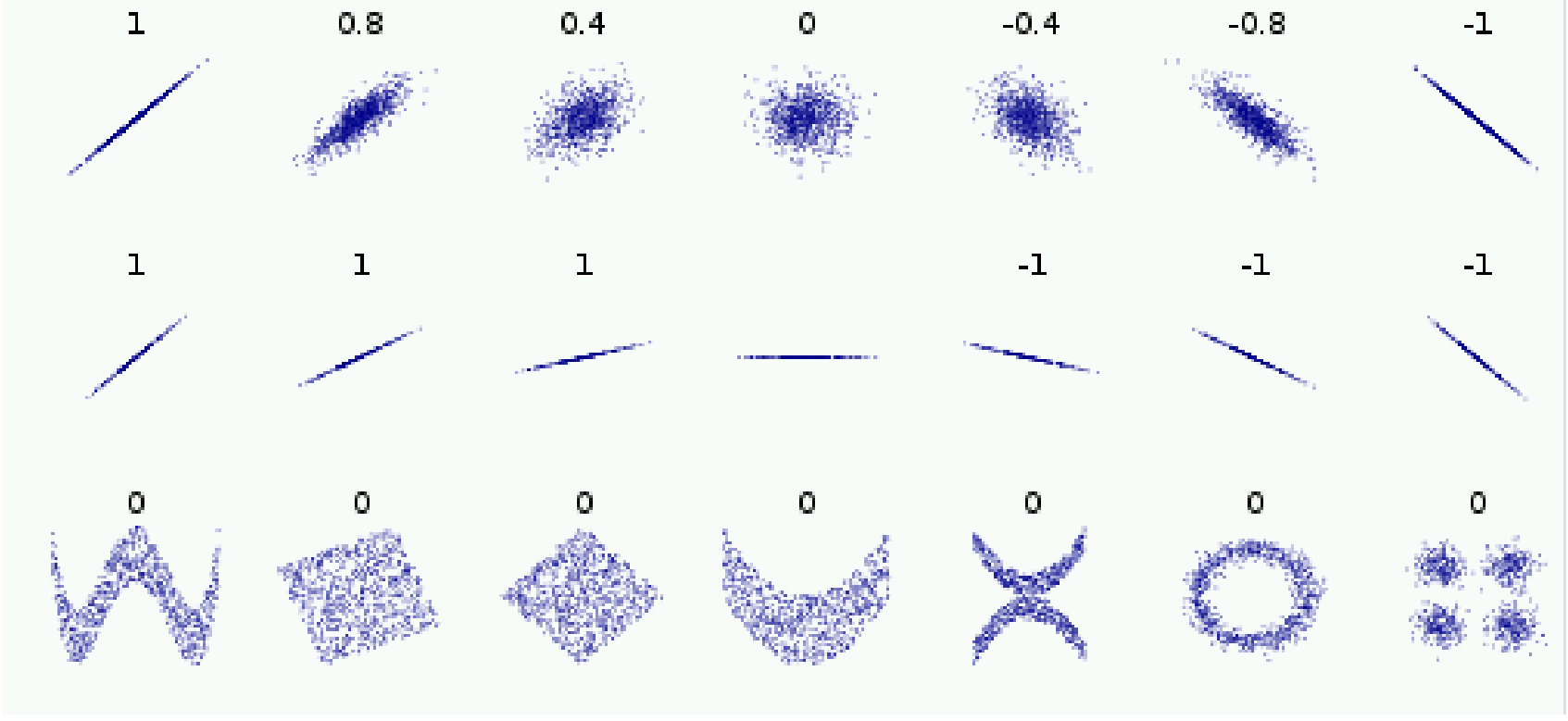


Figure 1: Correlation = LINEAR association

Random Vectors §2.5, 2.6

Suppose $\mathbf{X}$ is a $(p \times 1)$ random vector where each element is a r.v. with its own probability distribution.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \qquad E(\mathbf{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

Marginal Means:

$$\mu_i = E(X_i) = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i) dx_i & X_i \text{ is continuous r.v. with pdf } f_i(x_i) \\ \sum_{\text{all } x_i} x_i p_i(x_i) & X_i \text{ is discrete r.v. with pmf } p_i(x_i) \end{cases}$$

Marginal Variances:

$$\sigma_i^2 = E(X_i - \mu_i)^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i) dx_i \\ \sum_{\text{all } x_i} (x_i - \mu_i)^2 p_i(x_i) \end{cases}$$

Joint Distribution Function and Covariance

The behavior of any pair of random variables X_i and X_k is described by their joint prob function $f_{ik}(x_i, x_k)$ if both are cont and $p_{ik}(x_i, x_k)$ if both are discrete.

A measure of linear association between X_i and X_k is the covariance

$$\begin{aligned}
\sigma_{ik} &= Cov(X_i, X_k) \\
&= E(X_i - \mu_i)(X_k - \mu_k) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_k - \mu_k)f_{ik}(x_i, x_k)dx_i dx_k, \quad \text{if } X_i, X_k \text{ are cont} \\
&= \sum_{\text{all } x_i}\sum_{\text{all } x_k}(x_i - \mu_i)(x_k - \mu_k)p_{ik}(x_i, x_k), \quad \text{if } X_i, X_k \text{ are discrete}
\end{aligned}$$

Covariance Matrices

The population covariance matrix

$$\Sigma = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

1. If X_i and X_k are independent, then $\sigma_{ik} = 0$.
2. The converse is not true, in general. The converse holds for the multivariate normal.
3. The multivariate normal distribution is completely specified once the mean vector μ and variance-covariance matrix Σ are specified. More on this later.

Population Correllation Matrix

For $i \neq j$, the correlation between variables X_i and X_j is

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

Population correlation matrix and standard deviation matrix.

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \quad \mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Example: Multinomial Distribution

Let $\mathbf{X} = (X_1, X_2, X_3)$ follow the multinomial distribution with probability vector (p_1, p_2, p_3) with $p_1 + p_2 + p_3 = 1$.

$$
\begin{aligned}
f(\mathbf{x}) &= f(x_1, x_2, x_3) \\
&= \begin{cases} p_1^{x_1} p_2^{x_2} p_3^{x_3}, & \text{where } \mathbf{x} \in \{(1,0,0), (0,1,0), (0,0,1)\} \\ 0, & \text{elsewhere} \end{cases}
\end{aligned}
$$

1. Find the mean vector μ and variance-covariance matrix Σ.
2. Find the correlation matrix ρ.

Example 1.6 Lizard Size Data

```
# File "T1-3.dat"" should be in your working directory
# Specify the working directory
# Session > Set Working Directory > Choose Directory
lizard <- read.table("T1-3.dat")
# assign names to the column variables
colnames(lizard) <- c("mass", "svl", "hls")
head(lizard) # display first 6 rows
```

```
#     mass  svl   hls
# 1  5.526 59.0 113.5
# 2 10.401 75.0 142.0
# 3  9.213 69.0 124.0
# 4  8.953 67.5 125.0
# 5  7.063 62.0 129.5
# 6  6.610 62.0 123.0
```

Example 1.6 - Numerical Summary

```
colMeans(lizard) # Sample Mean Xbar of the three variables


#     mass      svl      hls
#   8.6866  68.4000 129.3200

# apply function computes summary by row or column
apply(lizard, MARGIN = 2, median) # median, MARGIN = 2 for


#    mass     svl     hls
#   8.953  68.000 129.500

apply(lizard, MARGIN = 2, var) # sample variance


#       mass        svl        hls
#   7.186551  63.770833 185.830833
```

Example 1.6 - Numerical Summary

```
cov(lizard)  # Sample Covariance matrix S
```

```
#              mass        svl        hls
# mass  7.186551  20.69952  33.53684
# svl  20.699521  63.77083 102.08542
# hls  33.536842 102.08542 185.83083
```

```
cor(lizard)  # Sample Correlation matrix R
```

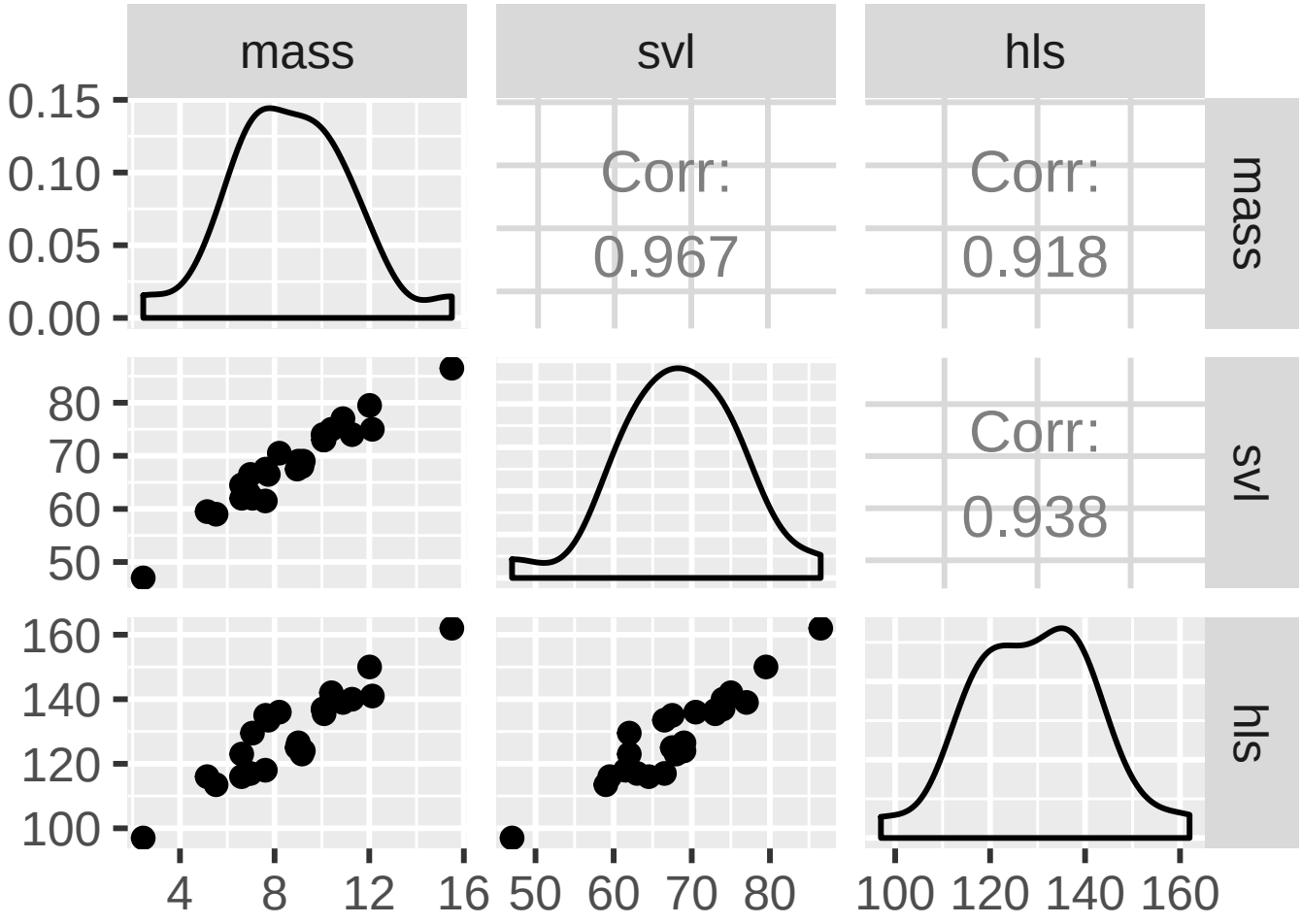
```
#              mass       svl       hls
# mass 1.0000000 0.9669165 0.9177048
# svl  0.9669165 1.0000000 0.9377645
# hls  0.9177048 0.9377645 1.0000000
```

What is the difference between the Sample Covariance Matrix and Sample Correlation Matrix?

Example 1.6 - Graphical Summary

Install the package `GGally` to use the function `ggpairs`

```
GGally::ggpairs(lizard)
```



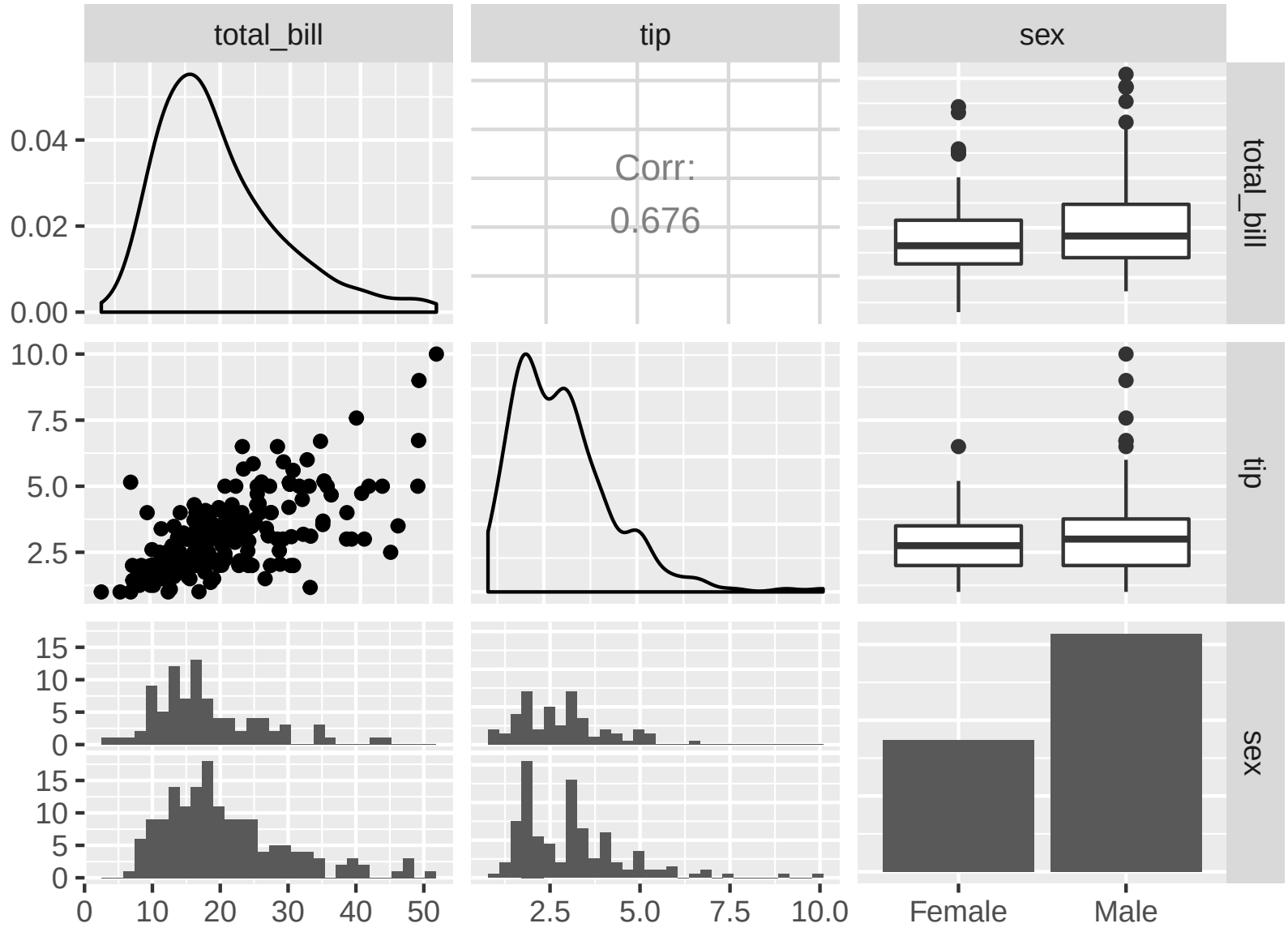
Load data from the web using its URL

One waiter recorded information about each tip he received over a period of a few months working in one restaurant.

```
tips <- read.csv(
  "http://siue.edu/~jpailde/tips.csv", header = TRUE)
str(tips) # check data structure
```

```
# 'data.frame': 244 obs. of  7 variables:
#  $ total_bill: num  17 10.3 21 23.7 24.6 ...
#  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1
#  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2
#  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1
#  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3
#  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1
#  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

```
GGally::ggpairs(tips[,1:3])
```



Properties of Expected Value

1. The linear combination

$$\mathbf{c}'\mathbf{X} = c_1 X_1 + \cdots + c_p X_p,$$

where c_i are constants, has

$$E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\mu$$
$$Var(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\Sigma_X\mathbf{c}$$

2. Let $\mathbf{C}$ be a $q \times p$ matrix of constants. The linear combinations $\mathbf{Z} = \mathbf{C}\mathbf{X}$ have

$$\mu_{\mathbf{Z}} = E(\mathbf{Z}) = E(\mathbf{C}\mathbf{X}) = \mathbf{C}\mu_{\mathbf{X}}$$
$$\Sigma_{\mathbf{Z}} = Cov(\mathbf{Z}) = Cov(\mathbf{C}\mathbf{X}) = \mathbf{C}\Sigma_X\mathbf{C}'$$

Properties of Expected Value

Result (1) is a special case of (2). To show (2), we need the following results:

Let $A_{n\times m}$, $B_{m\times n}$, $C_{m\times n}$.

Then,

$$A \cdot (B - C) = A \cdot B - A \cdot C$$

$$(A \cdot B)' = B' \cdot A'$$

Exercise 2.41

Let $\mathbf{X}' = [X_1, X_2, X_3, X_4]$ with mean vector $\mu'_X = [3, 2, -2, 0]$ and variance-covariance matrix

$$\text{(Case I) } \Sigma_X = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$\text{(Case II) } \Sigma_X = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}$$

Exercise 2.41

1. Under each cases, find $E(\mathbf{AX})$ and $Cov(\mathbf{AX})$ where

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -2 & 0 \\ 1 & 1 & 1 & -2 \end{bmatrix}$$

2. Let $\mathbf{X}^{(1)} = [X_1, X_2]'$ and $\mathbf{X}^{(2)} = [X_3, X_4]'$. Under Case II, find $Cov(\mathbf{AX}^{(1)}, \mathbf{BX}^{(\mathbf{2})})$ where

$$\mathbf{A} = [-1, 3] \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}$$